ON ATTENTION HEADS AND BILINEAR FORMS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the symmetric and antisymmetric parts of bilinear forms encoded by the query and key maps of attention heads, separating properties imposed by the architecture from patterns associated with training. We prove that, whenever the stacked query–key map is surjective, the symmetric part has exactly as many positive as negative eigenvalues. This condition is generic when the embedding dimension is at least twice the head dimension, and the predicted signature occurs in all $9{,}512$ heads of fourteen pretrained language models. To describe variation beyond this structural constraint, we introduce an orthogonally invariant profile map into a three-dimensional simplex. In thirteen of the fourteen models, trained profiles accumulate near the one-parameter family of profiles of rank-one forms. A limit theorem partially explains this phenomenon: profiles of symmetric matrices with proportionally distributed positive and negative eigenvalues converge to the projection of the rank-one profile family onto the symmetric face.

1 INTRODUCTION

An attention head compares input vectors using a bilinear form. If $Q, K : \mathbb{R}^N \to \mathbb{R}^n$ are its query and key maps, the bilinear term in its score is $x^T K^T Q y$, up to the standard factor $1/\sqrt{n}$ (Vaswani et al., 2017). This matrix admits a unique decomposition

$$L = K^T Q = S + T, \qquad S = \tfrac{1}{2}(L + L^T), \qquad T = \tfrac{1}{2}(L - L^T),$$

into symmetric and antisymmetric parts. The first gives scores unchanged by interchanging the inputs; the second changes sign. What properties of these forms distinguish trained attention heads, and which follow already from the query–key construction?

This distinction matters for interpreting eigenvalue measurements. Studying a self-suppressing GPT-2 head, Migliarini (2025) reported 33 negative eigenvalues among the 64 largest-magnitude eigenvalues of its symmetric part. In fact, the full nonzero spectrum has 64 positive and 64 negative eigenvalues. We prove that these equal counts follow from a generic rank condition on Q and K, rather than requiring an explanation in terms of the head's learned behavior. The same signature occurs in every head of the fourteen models we inspect.

Balance of counts still leaves substantial freedom in the eigenvalues. To describe that freedom together with the antisymmetric component, we assign each nonzero bilinear form a *profile* $\pi(L) = (a, b, c, d)$ in the simplex Δ_3. The coordinate a measures antisymmetric energy, b measures overlap between the positive and negative spectral lobes of S, and c, d measure their positive and negative discrepancies. The faces have explicit matrix descriptions (Figure 1). In particular, profiles of rank-one forms constitute two line segments, denoted Θ.

Our principal empirical finding is that profiles of trained bilinear forms accumulate near the profiles of rank-one forms in every surveyed model except `opt-350m` (Figure 3). We do not have a complete explanation for this. We give a partial explanation by proving a limit theorem for symmetric matrices whose positive and negative eigenvalues have proportional distributions. Their profiles converge to the two edges obtained by projecting Θ onto the symmetric face.

Contributions. We establish a generic balance theorem for attention heads (Section 2); define and characterize a profile map, including its rank-one locus (Section 3); and survey its behavior in $9{,}512$ heads spanning absolute, rotary, and ALiBi positional encodings (Section 4). Section 5 connects approximate proportionality of the spectral lobes to profile accumulation. Section 6 distinguishes

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

these conclusions from questions of mechanistic interpretation. Extended constructions and proofs appear in the appendices. In Appendix D we give a formulation of attention equivalent to the usual construction which we hope will be inviting to mathematicians interested in transformers.

Related work. Saponati et al. (2025) study the symmetric and antisymmetric components underlying self-attention, including symmetric bias; our experiments corroborate that bias in decoder-only models. Darveshi (2025) describe gamma-like distributions for absolute eigenvalues of Qwen3 attention matrices. We examine the two spectral lobes of S and the imaginary spectrum of T separately, compare gamma, lognormal, and Weibull fits, and relate their parameters.

2 Bilinear scores and structural balance

Let $C = \mathbb{R}^N$ and $H = \mathbb{R}^n$ carry their standard inner products. Throughout, $Q, K : C \to H$ are independent linear maps, and $\|\cdot\|$ denotes the Frobenius norm for matrices and the Euclidean norm for vectors. The signature of a real symmetric matrix is the triple (n_+, n_-, n_0) of its positive, negative, and zero eigenvalue counts. We call it *balanced* when $n_+ = n_-$. Appendix E recalls the necessary background on bilinear forms.

2.1 Positional encoding and the form being measured

QKV attention with additive positional vectors p_j has scores

$$s_{kj} = \frac{\langle K(x_j + p_j), Q(x_k + p_k)\rangle}{\sqrt{n}}, \qquad j \le k.$$

Their bilinear component is $x_j^T K^T Q x_k/\sqrt{n}$; the remaining terms are linear or constant in the inputs. With rotary position embeddings (Su et al., 2024), an orthogonal map $R : H \to H$ gives

$$s_{kj} = \frac{\langle Kx_j, R^{k-j} Q x_k\rangle}{\sqrt{n}}, \qquad L_d = K^T R^d Q.$$

Partial rotary embeddings act as the identity on the unrotated dimensions. ALiBi (Press et al., 2022) instead uses $s_{kj} = \langle Kx_j, Qx_k\rangle/\sqrt{n} - \mu(k-j)$ for a fixed positive scalar μ; its positional term is constant in the inputs. Each is a special case of *biaffine* scoring, meaning affine in each input separately. Appendix D gives the full formulation.

We measure $L = K^T Q$, omitting the positive factor $1/\sqrt{n}$, which changes neither signature nor profile. For the RoPE models, cross-model measurements concern the base form L_0, at relative position zero. The bilinear component excludes query/key biases and the lower-order terms introduced by additive positional embeddings or ALiBi.

2.2 The balance theorem

Call a head *nondegenerate* when its stacked map is surjective:

$$J = \begin{pmatrix} Q \\ K \end{pmatrix} : C \longrightarrow H \oplus H, \qquad Jx = (Qx, Kx).$$

Theorem 1. *If J is surjective, the symmetric part $S = \frac{1}{2}(K^T Q + Q^T K)$ has signature $(n, n, N - 2n)$. When $N \ge 2n$, the pairs (Q, K) satisfying this condition form a nonempty Zariski-open dense subset.*

In the language of bilinear forms, the symmetric form is the sum of a hyperbolic form with a zero form.

Proof. On $H \oplus H$, the symmetric form

$$B((q,k),(q',k')) = \tfrac{1}{2}\big(\langle k, q'\rangle + \langle q, k'\rangle\big)$$

has signature $(n, n, 0)$: it is positive definite on the diagonal subspace $\{(h, h)\}$ and negative definite on the anti-diagonal $\{(h, -h)\}$. Since $S(x, y) = B(Jx, Jy)$ and J is surjective, its radical is $\ker J$.

Restricting to any complement of $\ker J$ gives a form congruent to B, hence the stated signature. Surjectivity is the nonvanishing of some $2n \times 2n$ minor of J. It occurs, for example, when Q and K select disjoint sets of n coordinates, so this condition is nonempty and Zariski-open dense. □

Low rank alone is insufficient, for example $Q = K$ gives a positive semidefinite L. More generally, $\operatorname{rank} L \leq n$ implies only $n_+(S), n_-(S) \leq n$ (Appendix A). Nondegeneracy is what forces both counts to reach n. For each fixed RoPE distance d, replacing Q by $R^d Q$ preserves surjectivity, so the signature conclusion also holds for S_d.

The theorem concerns eigenvalue counts, not opposite eigenvalues of equal magnitude. On its nondegenerate quotient, S is *hyperbolic*: congruent to a sum of hyperbolic planes. A spectrum symmetric about zero is a stronger, inner-product-dependent property. Neither balance nor hyperbolicity alone specifies the spectral density.

3 Profiles of bilinear forms

Let λ_+ list the positive eigenvalues of S in decreasing order, and let λ_- list the absolute values of its negative eigenvalues in decreasing order. Append zeros to make the lengths equal. For a real scalar or vector, $(x)_+ = \max(x, 0)$ and $(x)_- = \max(-x, 0)$, entrywise for vectors.

Definition 1. For $L \neq 0$, its *profile* is $\pi(L) = (a, b, c, d)$, where

$$a = \frac{\|T\|^2}{\|L\|^2}, \qquad b = \frac{2\langle \lambda_+, \lambda_- \rangle}{\|L\|^2},$$
$$c = \frac{\|(\lambda_+ - \lambda_-)_+\|^2}{\|L\|^2}, \quad d = \frac{\|(\lambda_+ - \lambda_-)_-\|^2}{\|L\|^2}. \tag{1}$$

These coordinates are nonnegative and sum to one because $\|L\|^2 = \|S\|^2 + \|T\|^2$ and $\|S\|^2 = \|\lambda_+\|^2 + \|\lambda_-\|^2$. Thus π takes values in $\Delta_3 = \{(a, b, c, d) \in \mathbb{R}^4_{\geq 0} : a + b + c + d = 1\}$. It is invariant under positive scaling, orthogonal congruence, and adjoining zero rows and columns. Negating L exchanges c and d.

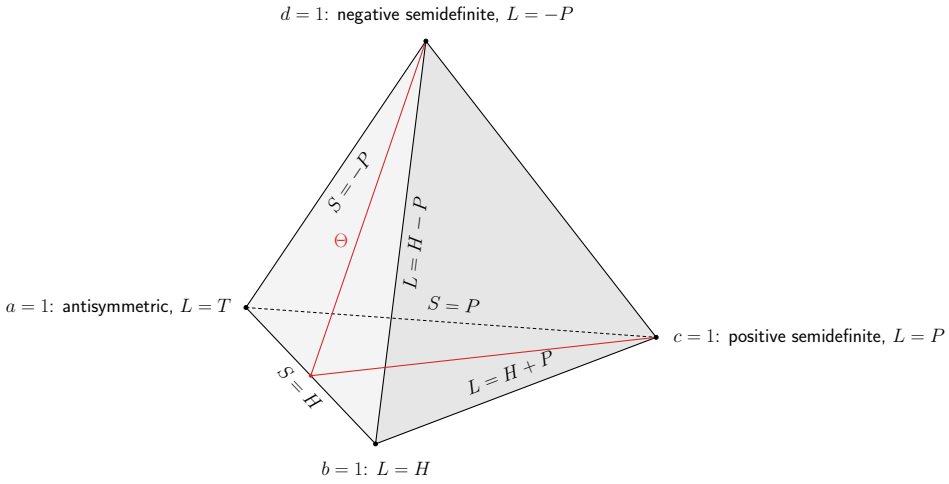

Figure 1: The profile simplex and the rank-one family Θ (red). Here H denotes a symmetric matrix with spectrum symmetric about zero and P a positive semidefinite matrix. The symmetric face is $a = 0$; $d = 0$ contains profiles of matrices for which $S = H + P$, and $c = 0$ those profiles for which $S = H - P$. The relative interior of $b = 0$ has empty preimage.

3.1 Faces and the rank-one family

The boundary of the simplex has a direct interpretation.

Theorem 2. *For $L = S + T \neq 0$, the following equivalences hold:*

1. *$a = 0$ if and only if L is symmetric; $a = 1$ if and only if L is antisymmetric.*
2. *$b = 0$ if and only if S is semidefinite; $b = 1$ if and only if L is symmetric with spectrum symmetric about zero.*
3. *$d = 0$ if and only if $S = H + P$, where H is symmetric with spectrum symmetric about zero and P is positive semidefinite. Equivalently, $(\lambda_+)_j \geq (\lambda_-)_j$ for every j.*
4. *$c = 0$ if and only if $S = H - P$ with H, P as above; equivalently, $(\lambda_-)_j \geq (\lambda_+)_j$ for every j.*

The vertices $c = 1$ and $d = 1$ correspond to symmetric positive and negative semidefinite forms, respectively.

The proof is in Appendix A.2. The profiles of rank-one matrices form a one-parameter subfamily $\Theta \subset \Delta_3$.

Proposition 1. *Let $N \geq 2$ and let $L \in M_N(\mathbb{R})$ have rank one and unit Frobenius norm. For $t = \operatorname{tr} L \in [-1, 1]$,*

$$\pi(L) = \left(\frac{1 - t^2}{2}, \frac{1 - t^2}{2}, (t)_+^2, (t)_-^2 \right). \tag{2}$$

The set of profiles of all rank-one matrices in $M_N(\mathbb{R})$ is $\Theta = \{a = b,\ cd = 0\} \subset \Delta_3$.

Indeed, writing $L = uv^T$ with u, v unit and $u^T v = t$, the nonzero eigenvalues of its symmetric part are $(1 + t)/2$ and $-(1 - t)/2$, and $\|T\|^2 = (1 - t^2)/2$. Every $t \in [-1, 1]$ is attainable in dimension at least two. The two segments meet at $\Theta_\circ = (1/2, 1/2, 0, 0)$ and end at the semidefinite vertices.

3.2 The symmetric profile and parity

We observed that eigenvalue distributions for the symmetric parts of trained attention heads are *bimodal*. They fall naturally into three types, called *parity*, depending on the relative position of the profile along Θ. For $S \neq 0$, write $\pi(S) = (0, b_S, c_S, d_S)$. A head is Type II when $c_S + d_S \leq 1/50$. Outside this neighborhood of the symmetric-spectrum vertex $(0, 1, 0, 0)$, the head is Type I_+ when $c_S > d_S$ and Type I_- otherwise. Type II consists of forms with symmetric spectrum with the stated tolerance; Types I_+ and I_- consist of balanced forms which are positive-definite-like or negative-definite-like, respectively.

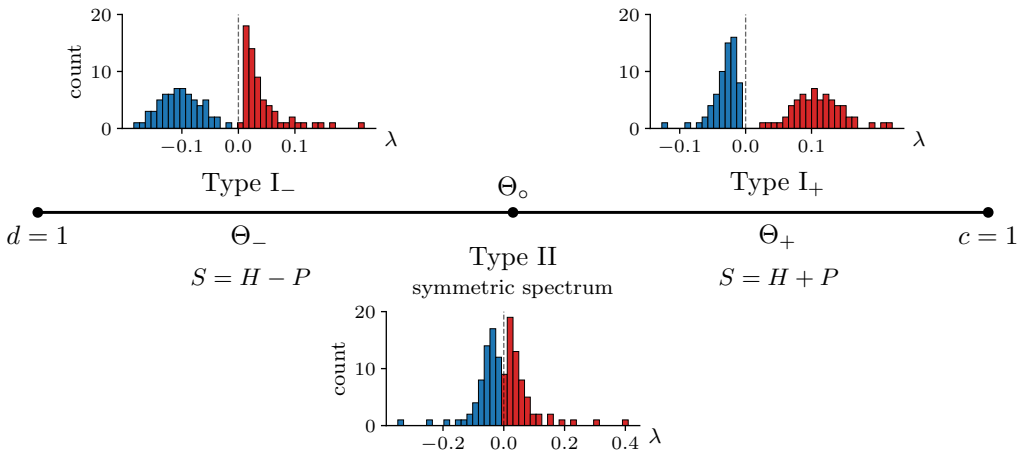

Figure 2: Measured eigenvalue distributions in `gpt2`, attached to their position along Θ. Each histogram shows the mean of the sorted, unit-length spectra of the ten most extreme heads in one cluster. Here H has spectrum symmetric about zero and P is positive semidefinite.

4 MEASUREMENTS ACROSS PRETRAINED MODELS

4.1 MODELS AND PROTOCOL

We analyze all heads of fourteen public decoder-only models: DistilGPT-2; GPT-2 small, medium, and large; OPT at 125M, 350M, 1.3B, 2.7B, 6.7B, and 13B; BLOOM at 3B and 7.1B; Pythia-70M-deduped; and Mistral-Small-24B-Base-2501 (Sanh et al., 2019; Radford et al., 2019; Zhang et al., 2022; BigScience Workshop et al., 2022; Biderman et al., 2023; Mistral AI Team, 2025). Together they contain 9,512 query heads with dimensions 64, 80, or 128. For grouped-query attention, each query head is paired with its associated shared key head. Appendix B lists dimensions, counts, and positional encodings.

We work on the joint row space of Q and K, of dimension at most $2n$, using an orthonormal basis. This preserves L, S, and T up to orthogonal congruence and zero padding, so it preserves their nonzero spectra, Frobenius norms, and profiles. Further details are in Appendix C.

Observation 1. All 9,512 heads have a symmetric part with exactly n positive and n negative eigenvalues, where n is the head dimension.

This agrees with Theorem 1. The observations below concern magnitudes and relative energies, which the theorem does not determine.

4.2 ACCUMULATION NEAR RANK-ONE PROFILES

Observation 2. In every surveyed model except `opt-350m`, the profiles of trained bilinear forms accumulate near Θ, the family of profiles of rank-one forms.

Figure 3 shows this pattern for GPT-2 large and Mistral Small, illustrating it for both additive and rotary positional encodings.[1] Gaussian initialization profiles lie tightly near $\Theta_\circ$; trained heads spread toward the two semidefinite vertices. Trained bilinear forms generally have large effective rank (Appendix B.2), so accumulation near Θ is not explained by proximity to low-rank forms.

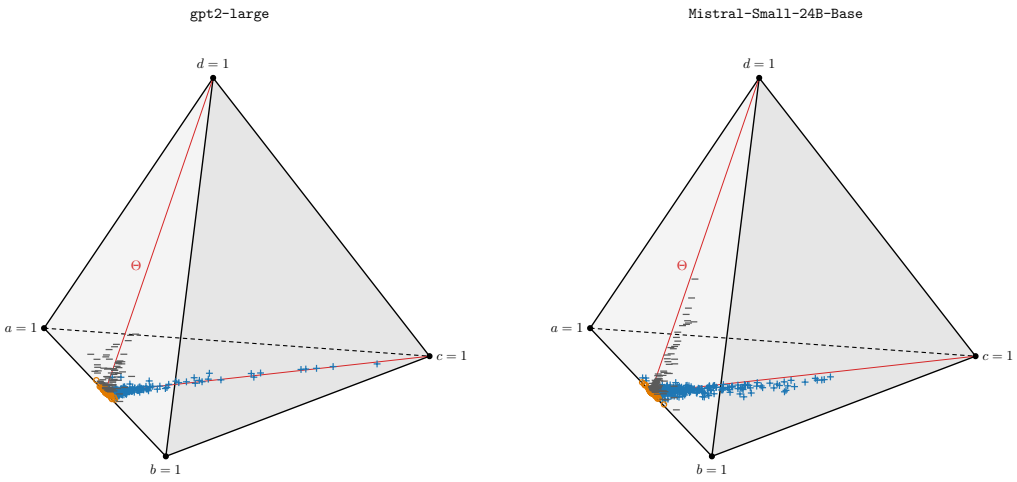

Figure 3: Profiles of all 720 heads of GPT-2 large (left) and all 1,280 query heads of Mistral Small (right; RoPE base form at relative position zero). Red segments are the rank-one family Θ. Blue, orange, and gray markers denote Types I_+, II, and I_-, respectively.

The head breakdowns by parity are 57.4% Type I_+, 28.5% Type II, and 14.1% Type I_-. Parity shows strong correlation with the head's layer position, though this correlation as well as parity proportions vary substantially across models (Table 2).

[1] The supplementary file `interactive/index.html` shows the profiles of all fourteen models in interactive plots.

4.3 SYMMETRIC BIAS AND PROPORTIONAL LOBES

Define the symmetric energy $\mathscr{S} = \|S\|^2/\|L\|^2 = 1 - a$. For independent Gaussian $n \times N$ query and key matrices with iid centered entries of positive variance,

$$\mathbb{E}[\mathscr{S}] = \frac{1}{2} + \frac{1}{2N}, \tag{3}$$

independently of n (Appendix A.3). Between 76% and 99% of the heads in each model exceed this baseline. This reproduces the symmetric bias reported by Saponati et al. (2025). It concerns energy, and is compatible with exactly balanced eigenvalue counts.

To compare the two sorted lobes, let $\rho = \|\lambda_-\|/\|\lambda_+\|$ and consider the relative discrepancy

$$\varepsilon = \frac{\|\lambda_- - \rho\lambda_+\|}{\|\lambda_-\|}.$$

Observation 3. The eigenvalue distribution of the symmetric part is bimodal, well-modelled by a sum of two gamma-like distributions, one supported on each side of zero (Figure 2); for 550 of the $9{,}512$ heads one of the two fitted modes is zero. The positive and negative lobes are approximately proportional after sorting: the discrepancy ε has median 0.13, with 80% of heads between 0.05 and 0.33.

The related pairing score

$$\mathscr{P} = 1 - \frac{\|\lambda_+ - \lambda_-\|}{\|S\|} \leq 1 - \frac{|1-\rho|}{\sqrt{1+\rho^2}} \tag{4}$$

attains equality precisely for proportional lobes. Per-model median gaps to this bound range from 0.017 to 0.041; median pairing scores range from 0.77 to 0.87. These observations motivate the proportional-distribution hypothesis in Section 5.

4.4 RELATIONS BETWEEN SYMMETRIC AND ANTISYMMETRIC SPECTRA

We fit gamma distributions to each positive spectral sample: λ_+, λ_-, the magnitudes of the nonzero eigenvalues of L, and the positive imaginary parts $\sigma(T)$ of the eigenvalues of T.

Gamma provides a descriptive parametrization but is not uniformly preferred across spectral quantities. We also fit lognormal and Weibull distributions. After normalizing each sample to unit mean, we compare the fitted density with a 20-bin histogram by root mean square error. Pooled median errors lie between 0.12 and 0.21 for every family and spectral quantity, with a spread between families of at most 0.072. Figure 4 gives the per-model comparisons.

Let $\mu \leq 0 \leq \nu$ denote the signed modes of the negative and positive gamma fits, and let k_-, k_+ be their shape parameters. Let τ and k_T be the mode and shape of the fit to $\sigma(T)$. We retain only heads with positive modes for all three magnitude distributions. This excludes 550 forms with at least one fitted mode zero and leaves $8{,}962$ heads.

Table 1: Relations between the symmetric and antisymmetric spectra. Each range is the minimum and maximum of the fourteen model medians. Here $s(T)$ is the full singular-value list on the reduced space, including the two copies corresponding to each conjugate eigenvalue pair. Correlations compare the two lists in decreasing order within a head.

Quantity	Per-model median	Heads included
$\tau/\sqrt{\lvert\mu\rvert\nu}$	1.01–1.05	8,962
$k_T/\sqrt{k_+k_-}$	1.02–1.15	8,962
Correlation of sorted $s(T)$ and $\lvert\lambda(S)\rvert$	0.986–0.997	9,512

We observed that the antisymmetric mode is close to the geometric mean of the two symmetric lobe modes, and the shape parameter has a similar relation. The sorted singular values of T and the sorted absolute eigenvalues of S are also highly correlated: per-model median correlations range from 0.986 to 0.997. These observations remain unexplained.

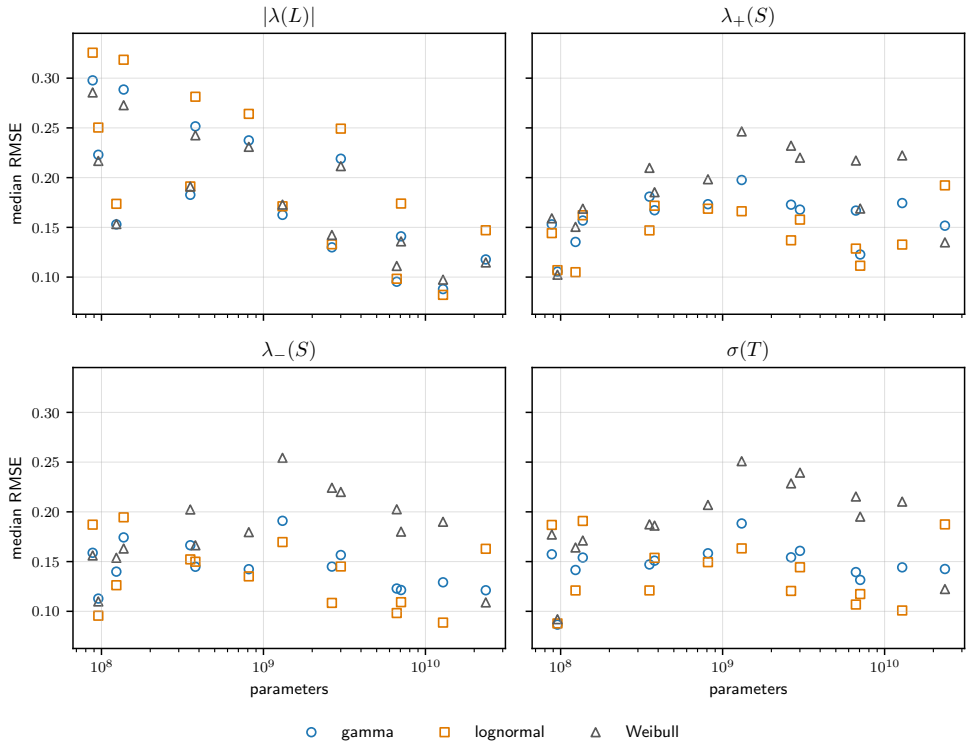

Figure 4: Comparison of gamma, lognormal, and Weibull fits to four spectral samples, plotted against model size. Each point gives the median histogram–density root mean square error over the heads of one model; each head's sample is first scaled to unit mean and binned into 20 equal-width bins. Lower is a closer fit. The gamma family is used as a descriptive parametrization.

Note that these relations hold exactly for rank-one forms. If $L = h\,uv^T$ for $h > 0$ with unit vectors u, v and $t = u^T v$, then S has the nonzero eigenvalues $h(1 \pm t)/2$ and T has the single repeated singular value $h\sqrt{1-t^2}/2$, their geometric mean. Nonetheless trained bilinear forms are far from rank one (Appendix B.2).

5 Why proportional lobes approach the rank-one family

Let $p : \Delta_3 \to \{a = 0\} : p(a, b, c, d) = (0, a + b, c, d)$ be the projection to the face of profiles of symmetric forms. Positive and negative eigenvalue distributions of the symmetric parts of trained bilinear forms are approximately proportional (Observation 3). The theorem below models this property and shows that the symmetric profiles converge to $p(\Theta)$ in the large-dimensional limit. This gives a partial explanation of Observation 2.

Theorem 3. *Let $X > 0$ almost surely and $0 < \mathbb{E}[X^2] < \infty$. Let $(X_i)_{i\geq 1}$ and $(X'_i)_{i\geq 1}$ be mutually independent iid sequences with the law of X. Fix $u, v > 0$ and put $\rho = u/v$. For each n, let $S_n \in M_{N_n}(\mathbb{R})$ be symmetric, with $N_n \geq 2n$, positive eigenvalues $vX_1, \ldots, vX_n$, negative eigenvalues $-uX'_1, \ldots, -uX'_n$, and all other eigenvalues zero. Then $\pi(S_n)$ converges almost surely to*

$$\frac{1}{1+\rho^2}\left(0, 2\rho, (1-\rho)_+^2, (\rho-1)_+^2\right) \in p(\Theta). \tag{5}$$

The proof is in Appendix A.4.

The limit depends on the ratio ρ, but not otherwise on the distribution of X. In particular, a gamma assumption is unnecessary. As ρ varies over $(0, \infty)$, the limits cover the two symmetric-face edges apart from their semidefinite endpoints; the endpoints are approached as $\rho \to 0$ or $\rho \to \infty$. For example, $\rho = 1/2$ gives $(0, 0.8, 0.2, 0)$, while $\rho = 1$ gives the symmetric-spectrum vertex $(0, 1, 0, 0)$.

6 Discussion and limitations

The query–key factorization imposes balanced inertia under a generic rank condition. Training alters how energy is divided between the symmetric and antisymmetric parts, how the two spectral lobes compare in magnitude, and where a head lies in the profile simplex. The profile map makes these distinctions visible and uncovers an empirical property of trained attention heads—accumulation near rank-one profiles—which remains to be fully explained.

The results concern fourteen decoder-only models. They do not establish a universal law across model families, training regimes, or inputs. The RoPE comparisons use relative position zero; distance-dependent profiles can vary even though the balance theorem continues to apply and accumulation along Θ continues to be observed. Gaussian random matrices give a mathematically controlled reference, rather than an exact account of every model's initialization. Head-level measurements also share weights and training history, so pooled summaries should not be interpreted as independent statistical replications.

Spectral profiles are coarse invariants, not complete mechanistic explanations. They depend on the chosen inner product and are not invariant under arbitrary invertible changes of residual-stream coordinates. They omit eigenvectors, value and output maps, nonlinear blocks, and the input distribution. Connecting profile types to specific components of linguistic processing remains future work.

The empirical scope is also limited. An attention head acts within a larger system: its output is combined with other heads and then processed by the neural network in each transformer block. Future work should examine how the symmetric and antisymmetric parts and their spectral distributions interact with these components.

The proportional-lobe theorem explains a limiting family on the symmetric face, while the full accumulation phenomenon remains open. The close relations between symmetric and antisymmetric spectral magnitudes and fitted parameters also remain unexplained. Distinguishing consequences of parametrization from effects of optimization and data is the next step toward understanding these patterns.

Reproducibility statement

The proof of the balance theorem is included in the main text; Appendix A supplies proofs for the profile geometry, random baseline, and limit theorem. Appendix B identifies the language models we surveyed, and Appendix C describes the extraction, normalization, parity partition, fitting procedures, and supplementary files. Theorem 2, Lemmas 2 and 3, and Proposition 2 have been formalized in Lean. The supplementary files include the formalization, experimental code and data.

AI use statement

Generative AI was used in developing the theoretical framework, formulating mathematical claims and proofs, designing and implementing experiments, interpreting results, preparing figures, and writing and editing the manuscript. Conversations with ChatGPT 5.5 contributed to the balance theorem and the attention constructions. ChatGPT 6 identified the rank-one profile family and assisted with mathematical review, experimental code, figures, and preparation of the manuscript. The proportional-lobe limit theorem and its proof arose from conversations with Claude Fable 5, which also generated experimental code, figures, and appendix material and assisted with Lean formalizations. The author is responsible for the final text, mathematical claims and proofs, empirical results, and artifacts.

References

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pp. 2397–2430. PMLR, 2023.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, et al. BLOOM: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022. doi: 10.48550/arXiv.2211.05100.

Shantanu Darveshi. Spectral taxonomy of QK circuits in transformer models. LessWrong, 2025. https://www.lesswrong.com/posts/Yig9fc7wAxKqG63Do/spectral-taxonomy-of-qk-circuits-in-transformer-models.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2nd edition, 2013.

Matteo Migliarini. The self-hating attention head: A deep dive in GPT-2. LessWrong / AI Alignment Forum, 2025. https://www.lesswrong.com/posts/wxPvdBwWeaneAsWRB/the-self-hating-attention-head-a-deep-dive-in-gpt-2-1.

Mistral AI Team. Mistral Small 3. https://mistral.ai/news/mistral-small-3/, January 2025.

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2108.12409.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Technical Report*, 2019.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Matteo Saponati, Pascal Sager, Pau Vilimelis Aceituno, Thilo Stadelmann, and Benjamin Grewe. The underlying structures of self-attention: symmetry, directionality, and emergent dynamics in transformer training. *arXiv preprint arXiv:2502.10927*, 2025.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

A PROOFS

A.1 RANK AND BALANCE

Lemma 1. *Let $L \in M_N(\mathbb{R})$ with $\operatorname{rank} L \le n$, and let (n_+, n_-, n_0) be the signature of its symmetric part $S = \frac{1}{2}(L + L^T)$. Then $n_+ \le n$ and $n_- \le n$.*

Proof. Let $K = \ker L$, of dimension at least $N - n$. For $x \in K$, $2\,x^T S x = x^T L x + x^T L^T x = 0$, so the quadratic form of S vanishes on K. Let V_+ be the span of the eigenvectors of S with positive eigenvalues, of dimension n_+. Then $V_+ \cap K = 0$: a nonzero vector x of the intersection would satisfy both $x^T S x > 0$ and $x^T S x = 0$. Hence $n_+ + (N - n) \le N$, that is, $n_+ \le n$; the same argument with the negative eigenvectors gives $n_- \le n$. □

A.2 FACES OF THE PROFILE SIMPLEX

Proof of Theorem 2. Rescaling, we may assume $\|L\| = 1$.

The partition. The entries of λ_+ and λ_- are the absolute values of the eigenvalues of S together with the appended zeros, so $\|\lambda_+\|^2 + \|\lambda_-\|^2 = \|S\|^2$; expanding $\|\lambda_+ - \lambda_-\|^2 = \|\lambda_+\|^2 + \|\lambda_-\|^2 - 2\langle \lambda_+, \lambda_- \rangle$ and using $\|\lambda_+ - \lambda_-\|^2 = \|(\lambda_+ - \lambda_-)_+\|^2 + \|(\lambda_+ - \lambda_-)_-\|^2$ gives

$$2\langle \lambda_+, \lambda_- \rangle + \|(\lambda_+ - \lambda_-)_+\|^2 + \|(\lambda_+ - \lambda_-)_-\|^2 = \|S\|^2.$$

Adding $a = \|T\|^2$ and using $\|S\|^2 + \|T\|^2 = \|L\|^2 = 1$ (Lemma 2) yields $a + b + c + d = 1$. Nonnegativity is clear since the entries of λ_+ and λ_- are nonnegative.

The facets. We prove the vanishing conditions of (1)–(4); the vertex conditions are proved below with the other zero-dimensional faces. If $S = 0$ then $b = c = d = 0$, S is semidefinite, and the decompositions of (3) and (4) hold with $H = P = 0$; so assume $S \neq 0$.

(1): $a = \|T\|^2$ vanishes if and only if $T = 0$.

(2): If S is semidefinite then $\lambda_+ = 0$ or $\lambda_- = 0$, so $b = 2\langle \lambda_+, \lambda_- \rangle = 0$. Conversely, if S is indefinite then $(\lambda_+)_1 > 0$ and $(\lambda_-)_1 > 0$; since the lists λ_+ and λ_- are sorted in decreasing order, $\langle \lambda_+, \lambda_- \rangle \ge (\lambda_+)_1 (\lambda_-)_1 > 0$, so $b > 0$.

(3): First, $d = \|(\lambda_+ - \lambda_-)_-\|^2$ vanishes iff $(\lambda_+)_j \ge (\lambda_-)_j$ for every j, which is the stated rank condition.

Assume the rank condition. Choose orthonormal eigenvectors of S: for each j with $(\lambda_+)_j > 0$ let v_j be an eigenvector for the eigenvalue $(\lambda_+)_j$, and for each j with $(\lambda_-)_j > 0$ let w_j be an eigenvector for $-(\lambda_-)_j$, taking each eigenvalue with its multiplicity. Since $(\lambda_-)_j > 0$ implies $(\lambda_+)_j > 0$, every w_j is matched with a v_j. Define

$$H = \sum_{j\,:\,(\lambda_-)_j > 0} (\lambda_-)_j \big(v_j v_j^T - w_j w_j^T \big), \qquad P = S - H = \sum_{j\,:\,(\lambda_+)_j > 0} \big((\lambda_+)_j - (\lambda_-)_j \big)\, v_j v_j^T.$$

The spectrum of H consists of the pairs $\pm(\lambda_-)_j$ together with zeros, so it is symmetric about zero; P is positive semidefinite since $(\lambda_+)_j \ge (\lambda_-)_j$; and $\|P\|^2 = \sum_j \big((\lambda_+)_j - (\lambda_-)_j \big)^2 = \|(\lambda_+ - \lambda_-)_+\|^2 = c$. All three matrices are diagonal in a common eigenbasis of S, so they commute.

Conversely, assume $S = H + P$. Let $\mu_1 \ge \cdots \ge \mu_N$ be the eigenvalues of H; symmetry of the spectrum gives $\mu_{N+1-k} = -\mu_k$. Writing $\lambda_k(S)$ for the k-th largest eigenvalue of S, the Courant–Fischer min–max theorem (Horn & Johnson, 2013, Theorem 4.2.6) gives

$$\lambda_k(S) = \max_{\dim V = k} \min_{\substack{x \in V \\ \|x\| = 1}} x^T (H + P)\, x \;\ge\; \max_{\dim V = k} \min_{\substack{x \in V \\ \|x\| = 1}} x^T H x = \mu_k,$$

since $x^T P x \ge 0$ for every x. Fix $t > 0$. If $\lambda_k(S) < -t$ then $\mu_k < -t$, so $\mu_{N+1-k} = -\mu_k > t$ and hence $\lambda_{N+1-k}(S) \ge \mu_{N+1-k} > t$. As $k \mapsto N + 1 - k$ is injective,

$$\#\{k : \lambda_k(S) < -t\} \;\le\; \#\{k : \lambda_k(S) > t\} \qquad \text{for every } t > 0. \tag{6}$$

Now suppose the rank condition fails, say $(\lambda_+)_j < (\lambda_-)_j$, and choose t with $(\lambda_+)_j < t < (\lambda_-)_j$. Since λ_+ is nonincreasing, the eigenvalues of S exceeding t are among $(\lambda_+)_i$ for $i < j$, so the right side of (6) is at most $j-1$; since $(\lambda_-)_i \geq (\lambda_-)_j > t$ for $i \leq j$, the left side is at least j. This contradicts (6).

(4): Apply (3) to $-L$, which preserves a and b, exchanges c with d and λ_+ with λ_-, and negates H and P.

The vertices. Each vertex is the face on which the other three coordinates vanish. If $a = 1$ then $c = d = 0$ gives $\lambda_+ = \lambda_-$ by (3) and (4), and $b = 0$ makes S semidefinite by (2), so $\lambda_+ = \lambda_- = 0$ and $S = 0$: L is antisymmetric. Conversely, an antisymmetric L has $S = 0$, so $a = \|T\|^2 = 1$. If $b = 1$ then L is symmetric with $\lambda_+ = \lambda_-$ by (1), (3), and (4); the converse is immediate from the definitions and the partition. If $c = 1$ then L is symmetric by (1) and S is semidefinite by (2); were S negative semidefinite, then $\lambda_+ = 0$ and $d = \|\lambda_-\|^2 = \|S\|^2 = 1$, contradicting $d = 0$; so L is positive semidefinite. Conversely, a symmetric positive semidefinite L has $T = 0$ and $\lambda_- = 0$, so $a = b = d = 0$ and $c = 1$. The case $d = 1$ follows by applying this to $-L$, which exchanges λ_+ with λ_- and hence c with d. □

A.3 The Gaussian reference

Lemma 2 (Symmetric and skew energy identities; cf. (Saponati et al., 2025, Lemma S1.14 and Eq. (S118))). *Let $L \in M_N(\mathbb{R})$, and write $S = \frac{1}{2}(L + L^T)$, $T = \frac{1}{2}(L - L^T)$. Then $\|L\|^2 = \|S\|^2 + \|T\|^2$ and $\|S\|^2 = \frac{1}{2}\|L\|^2 + \frac{1}{2}\operatorname{tr}(L^2)$, $\|T\|^2 = \frac{1}{2}\|L\|^2 - \frac{1}{2}\operatorname{tr}(L^2)$. Consequently,*

$$\frac{\|S\|^2}{\|L\|^2} = \frac{1}{2} + \frac{\operatorname{tr}(L^2)}{2\|L\|^2}, \qquad \frac{\|T\|^2}{\|L\|^2} = \frac{1}{2} - \frac{\operatorname{tr}(L^2)}{2\|L\|^2}.$$

Proof. The subspaces of symmetric and antisymmetric matrices are orthogonal with respect to the Frobenius inner product $\langle X, Y\rangle = \operatorname{tr}(XY^T)$: for S symmetric and T antisymmetric,

$$\langle S, T\rangle = \operatorname{tr}(ST^T) = -\operatorname{tr}(ST) = -\operatorname{tr}\big((ST)^T\big) = -\operatorname{tr}(T^T S^T) = \operatorname{tr}(TS) = \operatorname{tr}(ST),$$

so $\operatorname{tr}(ST) = 0$ and $\langle S, T\rangle = 0$. Since $L = S + T$, it follows that $\|L\|^2 = \|S\|^2 + \|T\|^2$. Expanding,

$$\|S\|^2 = \tfrac{1}{4}\big(\langle L, L\rangle + 2\langle L, L^T\rangle + \langle L^T, L^T\rangle\big).$$

Since $\langle L, L\rangle = \langle L^T, L^T\rangle = \|L\|^2$ and $\langle L, L^T\rangle = \operatorname{tr}(L^2)$, we obtain $\|S\|^2 = \frac{1}{2}\|L\|^2 + \frac{1}{2}\operatorname{tr}(L^2)$. The formula for $\|T\|^2$ follows by subtraction from $\|L\|^2$. □

Proposition 2 (Expected symmetric energy for random low-rank QK products). *Let $K, Q \in M_{n\times N}(\mathbb{R})$ be independent random matrices with iid centered Gaussian entries, where $1 \leq n \leq N$. Set $L = K^T Q \in M_N(\mathbb{R})$, $S = \frac{1}{2}(L + L^T)$, $T = \frac{1}{2}(L - L^T)$, and define*

$$\mathscr{S} = \frac{\|S\|^2}{\|L\|^2}, \qquad \mathscr{A} = \frac{\|T\|^2}{\|L\|^2}.$$

Then

$$\mathbb{E}[\mathscr{S}] = \frac{N+1}{2N}, \qquad \mathbb{E}[\mathscr{A}] = \frac{N-1}{2N},$$

independently of the head dimension n.

Proof. By Lemma 2,

$$\mathscr{S} = \frac{1}{2} + \frac{\operatorname{tr}(L^2)}{2\|L\|^2},$$

so it suffices to show that $\mathbb{E}[\operatorname{tr}(L^2)/\|L\|^2] = 1/N$.

Since K and Q have iid Gaussian entries, their joint law is invariant under right multiplication of either factor by any $V \in O_N(\mathbb{R})$. Therefore the law of $L = K^T Q$ is invariant under

$$L \longmapsto U^T L V$$

for every $U, V \in O_N(\mathbb{R})$. Set $X = L/\|L\|$, so $\|X\| = 1$ almost surely. By permutation invariance of the row and column indices, the expectations $\mathbb{E}[X_{ij}^2]$ are all equal, and since $\sum_{i,j} X_{ij}^2 = 1$,

$$\mathbb{E}[X_{ij}^2] = \frac{1}{N^2}$$

for every i, j. For $i \neq j$, changing the sign of the i-th column of K flips the i-th row of $L = K^T Q$, hence flips X_{ij} while leaving X_{ji} unchanged; this preserves the law of X, so $\mathbb{E}[X_{ij}X_{ji}] = 0$. Hence

$$\mathbb{E}\left[\frac{\mathrm{tr}(L^2)}{\|L\|^2}\right] = \mathbb{E}[\mathrm{tr}(X^2)] = \sum_{i,j} \mathbb{E}[X_{ij}X_{ji}] = \sum_{i=1}^{N} \mathbb{E}[X_{ii}^2] = \frac{1}{N},$$

and therefore $\mathbb{E}[\mathscr{S}] = \frac{1}{2} + \frac{1}{2N} = (N+1)/(2N)$. Since $\mathscr{A} = 1 - \mathscr{S}$, we get $\mathbb{E}[\mathscr{A}] = (N-1)/(2N)$. □

A.4 Proportional spectral lobes

Lemma 3. *Let X be a positive random variable with finite, nonzero second moment. Fix scales $u, v > 0$ and set $\rho = u/v$. For each n, let $\alpha_n \in \mathbb{R}^n$ be the decreasing sorting of $vX_1, \ldots, vX_n$ and let $\beta_n \in \mathbb{R}^n$ be the decreasing sorting of $uX_1', \ldots, uX_n'$, where the X_i and X_i' are independent copies of X. Then the sorted lobes become proportional with ratio ρ:*

$$\frac{\|\beta_n - \rho\,\alpha_n\|}{\|\alpha_n\|} \longrightarrow 0 \qquad \textit{almost surely as } n \to \infty.$$

Proof of Lemma 3. We may write $\beta_n = \rho\,\alpha_n'$, where α_n' is the decreasing sorting of $vX_1', \ldots, vX_n'$. Since the claim is unchanged when both lists are divided by v, we may assume $v = 1$. Let $m_2 = \mathbf{E}[X^2]$, which is finite and nonzero by assumption.

Let $\overline{F}(x) = \mathbf{P}(X > x)$, and let $G\colon (0,1) \to (0,\infty)$ be the decreasing generalized inverse $G(t) = \inf\{x > 0 : \overline{F}(x) \le t\}$, so that $G(t) > x$ if and only if $t < \overline{F}(x)$. Let U be a random variable with $\mathbf{P}(U < t) = t$ for $t \in [0,1]$. Then $\mathbf{P}(G(U) > x) = \mathbf{P}(U < \overline{F}(x)) = \overline{F}(x)$, so $G(U)$ has the law of X and

$$\int_0^1 G(t)^2\,dt = m_2.$$

Given n independent samples $X_1, \ldots, X_n$ of X with decreasing sorting $x_1 \ge \cdots \ge x_n$, define the step function $h_n(t) = x_{\lceil nt \rceil}$ on $(0,1)$, so that $\int_0^1 h_n^2 = \frac{1}{n}\sum_i x_i^2$. We claim that $h_n \to G$ in $L^2(0,1)$ almost surely, and prove it in three steps.

First, $h_n(t) \to G(t)$ almost surely for each fixed t at which G is continuous, hence for all but countably many $t \in (0,1)$. Let $\overline{F}_n(x) = n^{-1}\#\{i : X_i > x\}$, so that $h_n(t) > x$ exactly when $\overline{F}_n(x) \ge \lceil nt \rceil/n$. Fix $\varepsilon > 0$. Since G is continuous at t we have $\overline{F}(G(t)+\varepsilon) < t$, so by the strong law of large numbers $\overline{F}_n(G(t)+\varepsilon) \to \overline{F}(G(t)+\varepsilon) < t$ almost surely; eventually $\overline{F}_n(G(t)+\varepsilon) < t \le \lceil nt \rceil/n$ and hence $h_n(t) \le G(t) + \varepsilon$. The lower bound is symmetric, using $\overline{F}(G(t)-\varepsilon) > t$.

Second, by the strong law of large numbers,

$$\int_0^1 h_n(t)^2\,dt = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \longrightarrow m_2 = \int_0^1 G(t)^2\,dt \qquad \text{almost surely.}$$

Third, expanding $\int (h_n - G)^2 = \int h_n^2 + \int G^2 - 2\int h_n G$ and applying Fatou's lemma to the nonnegative functions $h_n G$, which converge to G^2 almost everywhere by the first step and so give $\liminf \int h_n G \ge \int G^2$, we conclude

$$\limsup_n \int_0^1 (h_n - G)^2\,dt \;\le\; m_2 + m_2 - 2m_2 = 0,$$

which proves the claim.

Now let g_n and g'_n be the step functions of the two samples defining α_n and α'_n. Then

$$\frac{\|\alpha_n - \alpha'_n\|^2}{n} = \int_0^1 (g_n - g'_n)^2$$
$$\leq 2\int_0^1 (g_n - G)^2 + 2\int_0^1 (g'_n - G)^2 \longrightarrow 0,$$
$$\frac{\|\alpha_n\|^2}{n} = \int_0^1 g_n^2 \longrightarrow m_2 > 0,$$

almost surely, and therefore $\|\beta_n - \rho\,\alpha_n\|/\|\alpha_n\| = \rho\,\|\alpha'_n - \alpha_n\|/\|\alpha_n\| \to 0$. □

Proof of Theorem 3. Let α_n and β_n denote the positive eigenvalues of S_n and the absolute values of its negative eigenvalues, each listed in decreasing order, so that $\lambda_+ = \alpha_n$ and $\lambda_- = \beta_n$. By Lemma 3 the sorted lobes become proportional in the limit: $\|\beta_n - \rho\,\alpha_n\|/\|\alpha_n\| \to 0$ almost surely. Write $\beta_n = \rho\,\alpha_n + \varepsilon_n$ with $\delta_n = \|\varepsilon_n\|/\|\alpha_n\| \to 0$. By the Cauchy–Schwarz and triangle inequalities,

$$\Big|\frac{\langle \alpha_n, \beta_n\rangle}{\|\alpha_n\|^2} - \rho\Big| \leq \delta_n, \qquad \Big|\frac{\|\beta_n\|^2}{\|\alpha_n\|^2} - \rho^2\Big| \leq \delta_n^2 + 2\rho\,\delta_n,$$

and, since $x \mapsto x_\pm$ is 1-Lipschitz in each entry and $\big((1-\rho)\,\alpha_n\big)_\pm = (1-\rho)_\pm\,\alpha_n$,

$$\Big|\frac{\|(\alpha_n - \beta_n)_\pm\|}{\|\alpha_n\|} - (1-\rho)_\pm\Big| \leq \frac{\|(\alpha_n - \beta_n) - (1-\rho)\,\alpha_n\|}{\|\alpha_n\|} = \delta_n.$$

Since $\|S_n\|^2 = \|\alpha_n\|^2 + \|\beta_n\|^2$, dividing each of $\|S_n\|^2$, $2\langle\alpha_n, \beta_n\rangle$ and $\|(\alpha_n - \beta_n)_\pm\|^2$ by $\|\alpha_n\|^2$ gives quantities that converge almost surely to $1+\rho^2$, 2ρ and $(1-\rho)_\pm^2$ respectively. The profile map is homogeneous of degree zero in S_n, so

$$b = \frac{2\langle\alpha_n, \beta_n\rangle}{\|S_n\|^2} = \frac{2\langle\alpha_n, \beta_n\rangle/\|\alpha_n\|^2}{\|S_n\|^2/\|\alpha_n\|^2},$$

and likewise for c and d; the factor $\|\alpha_n\|^2$ cancels, and each ratio converges almost surely to the corresponding entry of (5). □

B MODELS

We analyze fourteen publicly available pretrained language models, comprising 9,512 attention heads. The models have head dimensions 64, 80, or 128 and use learned absolute positional embeddings, ALiBi, or rotary position embeddings (RoPE):

- `distilgpt2` Sanh et al. (2019): distilled `gpt2`, 6 layers, 12 heads per layer (72 heads), $N = 768$, and $n = 64$.
- `gpt2` Radford et al. (2019): GPT-2 small, 12 layers, 12 heads per layer (144 heads), $N = 768$, and $n = 64$.
- `gpt2-medium` Radford et al. (2019): 24 layers, 16 heads per layer (384 heads), $N = 1024$, and $n = 64$.
- `gpt2-large` Radford et al. (2019): 36 layers, 20 heads per layer (720 heads), $N = 1280$, and $n = 64$.
- `opt-125m` Zhang et al. (2022): OPT decoder-only, 12 layers, 12 heads per layer (144 heads), $N = 768$, and $n = 64$.
- `opt-350m` Zhang et al. (2022): 24 layers, 16 heads per layer (384 heads), $N = 1024$, and $n = 64$.
- `opt-1.3b` Zhang et al. (2022): 24 layers, 32 heads per layer (768 heads), $N = 2048$, and $n = 64$.
- `opt-2.7b` Zhang et al. (2022): 32 layers, 32 heads per layer (1,024 heads), $N = 2560$, and $n = 80$.

- `opt-6.7b` Zhang et al. (2022): 32 layers, 32 heads per layer (1,024 heads), $N = 4096$, and $n = 128$.
- `opt-13b` Zhang et al. (2022): 40 layers, 40 heads per layer (1,600 heads), $N = 5120$, and $n = 128$.
- `bloom-3b` BigScience Workshop et al. (2022): 30 layers, 32 heads per layer (960 heads), $N = 2560$, $n = 80$, and ALiBi positional encoding.
- `bloom-7b1` BigScience Workshop et al. (2022): 30 layers, 32 heads per layer (960 heads), $N = 4096$, $n = 128$, and ALiBi positional encoding.
- `pythia-70m-deduped` Biderman et al. (2023): GPT-NeoX family, 6 layers, 8 heads per layer (48 heads), $N = 512$, $n = 64$, and RoPE on the first 25% of each head's dimensions.
- `Mistral-Small-24B-Base-2501` Mistral AI Team (2025): 40 layers, 32 query heads per layer (1,280 heads), 8 key–value heads, $N = 5120$, $n = 128$, and RoPE.

B.1 The parity of trained heads

Table 2: Percentage of the heads by parity, classified using $\pi(S)$, with tolerance $c_S + d_S \leq \frac{1}{50}$ for Type II.

model	heads	Type I_+	Type II	Type I_-
`distilgpt2`	72	34.7	26.4	38.9
`gpt2`	144	36.8	34.0	29.2
`gpt2-medium`	384	54.2	28.6	17.2
`gpt2-large`	720	64.3	22.6	13.1
`opt-125m`	144	79.2	16.7	4.2
`opt-350m`	384	78.9	20.8	0.3
`opt-1.3b`	768	72.3	17.6	10.2
`opt-2.7b`	1,024	71.4	14.8	13.8
`opt-6.7b`	1,024	78.7	15.9	5.4
`opt-13b`	1,600	75.7	6.0	18.3
`bloom-3b`	960	29.7	49.9	20.4
`bloom-7b1`	960	23.2	55.2	21.6
`pythia-70m-deduped`	48	18.8	50.0	31.2
`Mistral-Small-24B-Base`	1,280	36.8	53.8	9.5
all	9,512	57.4	28.5	14.1

B.2 Effective rank of trained bilinear forms

Let $\sigma_1 \geq \sigma_2 \geq \cdots$ be the singular values of L. One measure of effective rank is the *participation ratio* $(\sum_i \sigma_i^2)^2 / \sum_i \sigma_i^4$, which equals 1 for a rank-one form and n for a flat spectrum. Over the 9,512 heads it has median 48.5 and lies below 2 for fewer than 1% of them (`scripts/effective_rank.py`).

C Experimental conventions

A QKV attention head A is associated with a single bilinear form B_A on C^2. In the notation of Vaswani et al. (2017) where $C = \mathbb{R}^N$ and $H = \mathbb{R}^n$ and $Q, K \in M_{n\times N}(\mathbb{R})$, this bilinear form is given by $B_A(x, y) = x^T K^T Q y$. We write L_A for the Gram matrix $K^T Q$. For `pythia-70m-deduped` and `Mistral-Small-24B-Base-2501`, which use RoPE attention, the form $L_A = K^T Q$ reported in the cross-model comparisons is the base form at relative position $d = 0$.

C.1 Extraction, normalization, and parity

For each head, we form an orthonormal basis R of the joint row space by a reduced QR decomposition of $[K^T\ Q^T]$, and store $R^T L_A R = (KR)^T(QR)$. Grouped-query heads use their associated shared key head. The reduced matrices are stored in float32; spectral calculations use float64. For eigenvalue counts, an eigenvalue is treated as zero when its absolute value is at most 10^{-9} times the largest absolute eigenvalue of the symmetric part. The factor $1/\sqrt{n}$ and query/key biases are omitted, as in Section 2.

Profiles use the squared Frobenius norm of L for normalization; symmetric profiles use that of S. The positive lobe and the absolute negative lobe are sorted in decreasing order and zero-padded to equal length. Table 2, the profile scatter plots, and Figure 9 use the same partition: Type II has $c_S + d_S \leq 1/50$; otherwise $c_S > d_S$ gives Type I_+, and the remaining heads are Type I_-. Figure 2 uses a separate illustrative three-cluster selection, with the sorted spectrum of each selected head scaled to unit Euclidean norm before averaging.

C.2 Gamma vs. lognormal vs. Weibull

We compared the gamma distribution's fit with two other families. For each head we fitted a gamma, a lognormal and a Weibull distribution to each of the following quantities: the magnitudes $|\lambda(L)|$ of the nonzero eigenvalues of the full form, the two lobes $\lambda_+(S)$ and $\lambda_-(S)$ of the spectrum of the symmetric part, and the positive imaginary parts $\sigma(T)$ of the eigenvalues of the antisymmetric part. Fits use maximum likelihood with location fixed at zero. To compare errors, each spectral sample is scaled to unit mean and fitted again; we compute the root mean square difference between a 20-bin density histogram and the fitted density evaluated at the bin midpoints. Figure 4 reports the median of this error over each model's heads, separately for each family and spectral quantity. Table 1 uses the full singular-value list of T, including multiplicities, for correlations, and excludes heads with a zero fitted mode only from the mode and shape ratios.

C.3 Supplementary files

The anonymous supplement contains the Lean project in `formalization/`, experimental entry points in `scripts/`, and cached spectra, statistics, and figures. The three numbered observation scripts follow the numbering of this paper. The root `README.md` maps results and figures to scripts and gives build and reproduction commands; the formalization README lists the precise Lean coverage. The default reproduction uses cached data without downloading model weights. An optional matrix archive supports recomputing the spectra and statistics; exact checkpoint revisions and head assignments are recorded in the supplied SQLite database.

D Biaffine attention and transformers

We give a mathematical formulation of transformers which is equivalent to the usual construction. Let C be a finite-dimensional real vector space and fix a sequence length ℓ. We describe causal decoder-only transformers, omitting layer normalization. The construction proceeds from attention functions on prefixes to attention heads, multi-head blocks, and residual transformer blocks. Write $\Delta_{k-1} = \{\omega \in \mathbb{R}^k_{\geq 0} : \sum_{j=1}^k \omega_j = 1\}$.

D.1 Attention functions

First we define a very general class of attention functions.

Definition 2. A smooth function $a\colon C^k \to C$ is an *attention function* if there are a smooth function

$$\omega = (\omega_1, \ldots, \omega_k)\colon C^k \to \Delta_{k-1},$$

a smooth convex function $\phi\colon \mathbb{R}^k \to \mathbb{R}$ and a polynomial map $\sigma\colon C^k \to \mathbb{R}^k$ such that

$$\omega = (\nabla\phi) \circ \sigma \qquad \text{and} \qquad a(x_1, \ldots, x_k) = \sum_{j=1}^{k} \omega_j(x_1, \ldots, x_k) x_j.$$

The components of ω are called the *attention weights*, σ is the *scoring function*, and ϕ we call the *potential function*; with an abuse of terminology, we regard ω, σ, ϕ as part of the data of the attention function.

The usual choice is the *softmax potential*

$$\phi(s_1, \dots, s_k) = \log\left(\sum_{j=1}^{k} e^{s_j}\right), \qquad \omega_j(x) = \frac{e^{\sigma_j(x)}}{\sum_{i=1}^{k} e^{\sigma_i(x)}}.$$

Its gradient takes values in Δ_{k-1}, so the attention function returns a convex combination of its inputs.

D.2 Biaffine scores

Definition 3. Let $a\colon C^k \to C$ be an attention function with scoring function $\sigma\colon C^k \to \mathbb{R}^k$ of the form

$$\sigma(x_1, \dots, x_k) = \big(P_{k-1}(x_1, x_k), P_{k-2}(x_2, x_k), \dots, P_1(x_{k-1}, x_k), P_0(x_k, x_k)\big)$$

for polynomials $P_j\colon C^2 \to \mathbb{R}$. If the polynomials P_j all have bidegree at most $(1,1)$ then we say a is *biaffine*. If the biaffine polynomials are bilinear forms then we say that a is *bilinear*.

The following elementary lemma shows that the bilinear, linear, and constant components of a biaffine score are well-defined; we use it to attach a bilinear form to each attention head in §4.

Lemma 4. *Every polynomial $P : C \times C \to \mathbb{R}$ of bidegree at most $(1,1)$ admits a unique decomposition*

$$P(u, v) = B(u, v) + L(u) + R(v) + D,$$

where B is bilinear, L and R are linear, and $D \in \mathbb{R}$. The components are given by

$$D = P(0,0), \;\; L(u) = P(u,0) - P(0,0), \;\; R(v) = P(0,v) - P(0,0),$$

and $B(u,v) = P(u,v) - P(u,0) - P(0,v) + P(0,0)$.

D.3 Causal attention heads

Definition 4. For each $k = 2, \dots, \ell$, let $a_k\colon C^k \to C$ be an attention function. The associated *(causal) attention head* is

$$A(x_1, \dots, x_\ell) = \big(x_1, a_2(x_1, x_2), \dots, a_\ell(x_1, \dots, x_\ell)\big).$$

The head is *biaffine* or *bilinear* when every a_k has the corresponding type of scoring function.

The k-th output depends only on the prefix $(x_1, \dots, x_k)$; this is the causal restriction. Writing its weights as $\omega_{kj}(x)$, with $\omega_{11} = 1$, gives

$$A(x)_k = \sum_{j=1}^{k} \omega_{kj}(x) x_j, \qquad \omega_{kj}(x) \geq 0, \qquad \sum_{j=1}^{k} \omega_{kj}(x) = 1.$$

The scoring polynomials may depend on both positions j, k, as in QKV attention with fixed positional embeddings. QKV, RoPE, and ALiBi heads arise by applying the softmax potential to the respective score polynomials in Section 2.

D.4 Multi-head attention blocks

Let $A_1, \dots, A_h\colon C^\ell \to C^\ell$ be attention heads, and let $V_1, \dots, V_h\colon C \to C$ be linear maps. The associated *attention block* $A\colon C^\ell \to C^\ell$ is defined by

$$A(x)_k = \sum_{i=1}^{h} V_i\big(A_i(x)_k\big), \qquad 1 \leq k \leq \ell.$$

Thus each head first averages inputs using its own scores, applies its value map, and contributes to the sum.

In the standard multi-head construction (Vaswani et al., 2017, §3.2.2), choose head spaces H_i and query and key maps $Q_i, K_i \colon C \to H_i$. The combined value and output map factors as

$$V_i = W_{O,i} W_{V,i}, \qquad W_{V,i} \colon C \to H_i, \qquad W_{O,i} \colon H_i \to C.$$

By linearity, the resulting block has the familiar form

$$A(x)_k = \sum_{i=1}^{h} W_{O,i} \left(\sum_{j=1}^{k} \omega_{kj}^{(i)}(x) W_{V,i} x_j \right).$$

This is also obtained by concatenating the head outputs and applying $W_O \colon \bigoplus_i H_i \to C$, $W_O(z_1, \dots, z_h) = \sum_i W_{O,i} z_i$.

D.5 Transformers

Let $A \colon C^\ell \to C^\ell$ be an attention block and let $M \colon C \to C$ be a piecewise-smooth neural network. Define the residual attention and neural-network components by

$$T_A(x) = x + A(x),$$
$$T_M(y_1, \dots, y_\ell) = \big(y_1 + M(y_1), \dots, y_\ell + M(y_\ell)\big).$$

For example, M may have the form $M(y) = W_2 \rho(W_1 y)$, where ρ is a coordinatewise activation such as ReLU.

Definition 5. The *transformer block* associated with A, M is $T = T_M \circ T_A \colon C^\ell \to C^\ell$. Explicitly,

$$T(x)_k = x_k + A(x)_k + M\big(x_k + A(x)_k\big).$$

The network M acts separately at each position; interaction between positions comes from the attention block. A transformer is called biaffine or bilinear when its heads have the corresponding type of scores.

E Background on bilinear forms

Let C be an N-dimensional real vector space. A *bilinear form* is a map $B \colon C \times C \to \mathbb{R}$ that is linear in each argument separately. In a basis $e_1, \dots, e_N$, its *Gram matrix* is $L = (B(e_i, e_j))_{ij}$, so $B(x, y) = x^T L y$ for coordinate vectors x, y. A change of basis with invertible matrix P replaces L by $P^T L P$; this relation between matrices is called *congruence*.

The form is *symmetric* if $B(x, y) = B(y, x)$ and *antisymmetric* (or *skew-symmetric*) if $B(x, y) = -B(y, x)$. Every bilinear form has a unique decomposition into symmetric and antisymmetric parts. Using the same notation for forms and their matrices, this is

$$L = S + T, \qquad S = \tfrac{1}{2}(L + L^T), \qquad T = \tfrac{1}{2}(L - L^T).$$

For a symmetric or antisymmetric form, the *radical* is $\operatorname{rad} B = \{x \in C : B(x, y) = 0 \text{ for all } y \in C\}$, represented in coordinates by $\ker L$. The form is *nondegenerate* when this subspace is zero.

Sylvester's law of inertia states that a real symmetric form is determined up to congruence by its *signature* (n_+, n_-, n_0), the numbers of positive, negative, and zero eigenvalues of its Gram matrix, counted with multiplicity. These counts are independent of the basis: some invertible P gives

$$P^T S P = \operatorname{diag}(I_{n_+}, -I_{n_-}, 0_{n_0}).$$

The form is *positive semidefinite* when $S(x, x) \geq 0$ for all x, and *negative semidefinite* when $-S$ is positive semidefinite. For symmetric forms, $S_1 \succeq S_2$ means that $S_1 - S_2$ is positive semidefinite. A form is *balanced* when $n_+ = n_-$ and *hyperbolic* when it is also nondegenerate. Balance concerns eigenvalue counts; it does not require positive and negative eigenvalues to have equal magnitudes.

With an inner product fixed, orthonormal changes of basis preserve eigenvalues and Frobenius norms. The eigenvalues of S are real, while the nonzero eigenvalues of T occur in pairs $\pm i\sigma_j$, with $\sigma_j > 0$. The two parts are orthogonal for the Frobenius inner product, giving

$$\|L\|^2 = \|S\|^2 + \|T\|^2.$$

Thus the signature is independent of any inner product, whereas the spectral magnitudes and profiles used in this paper refer to the fixed Euclidean structure on the embedding space.

F SUPPLEMENTARY FIGURES

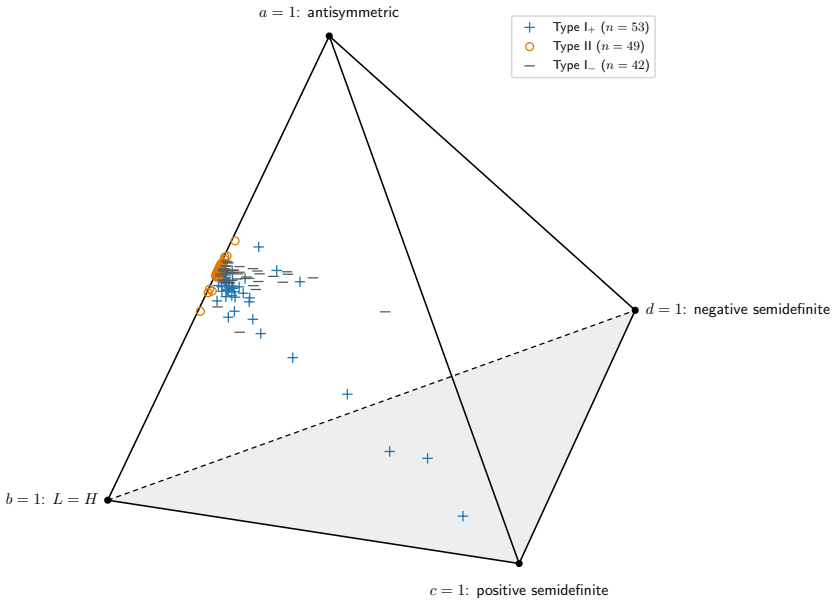

Figure 5: Profiles of the heads of `gpt2` on the simplex Δ_3.

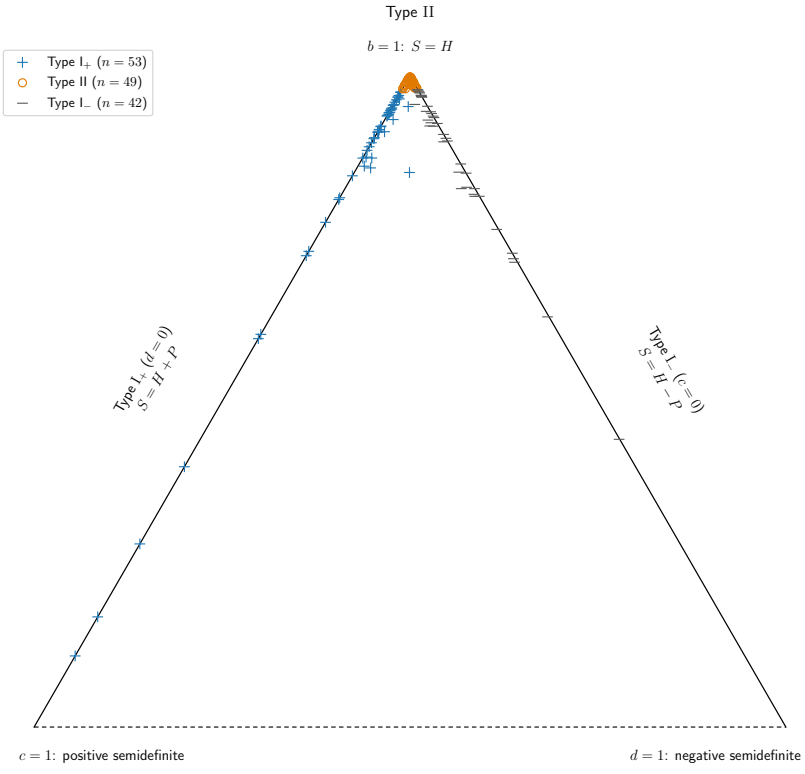

Figure 6: Profiles of the symmetric parts of the heads of `gpt2` on $p(\Delta_3) = \{a = 0\}$. Here H has spectrum symmetric about zero and P is positive semidefinite.

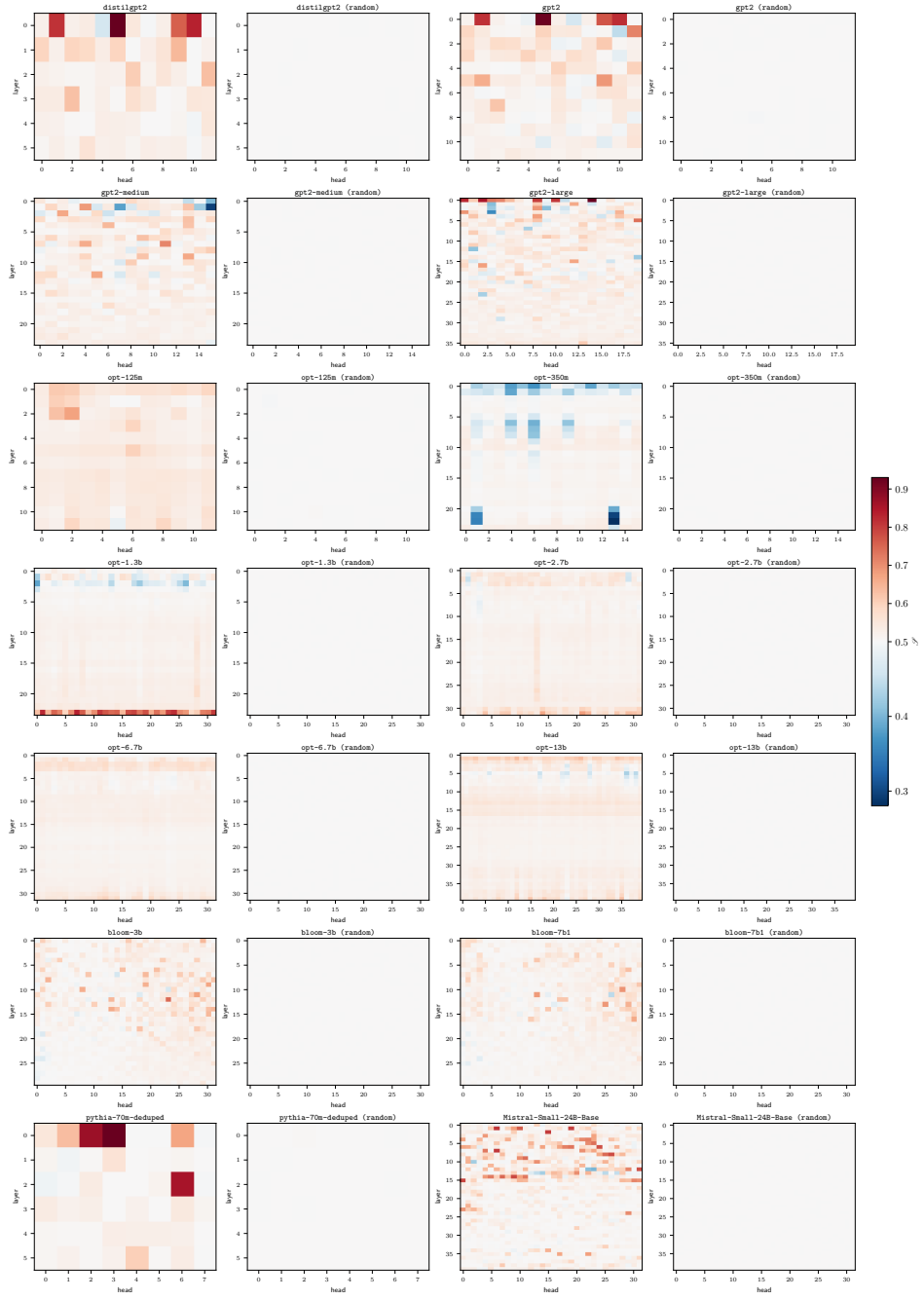

Figure 7: Symmetric energy $\mathscr{S} = \|S\|^2/\|L\|^2$ at each attention head. For each model, the trained heads (left) are paired with independent random products $L = K^T Q$ (right), with iid standard normal $K, Q \in \mathbb{R}^{n\times N}$ of the same dimensions. Rows index layers and columns index heads, both starting at zero. The shared color scale is white at $1/2$, red above and blue below; the random baseline has expectation $(N+1)/(2N)$. RoPE models use relative position $d = 0$.

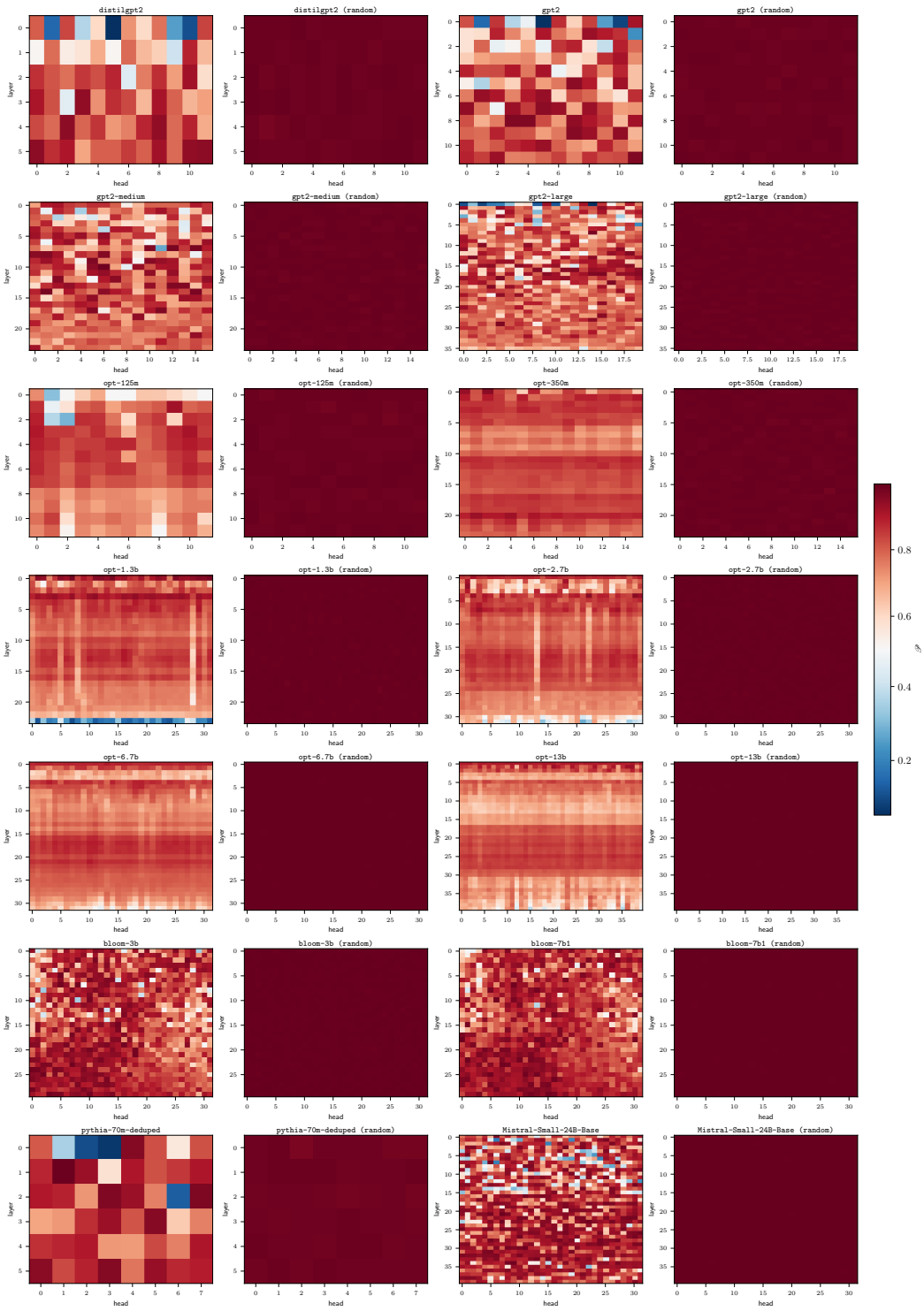

Figure 8: Pairing score $\mathscr{P} = 1 - \|\lambda_+ - \lambda_-\|/\|S\|$ at each attention head. Each model's trained heads (left) are paired with independent Gaussian QK products of the same dimensions (right), as in Figure 7. Rows index layers and columns index heads. A score of one means the symmetric spectrum is symmetric about zero. The shared color scale is white at $1/2$, red above and blue below. RoPE models use relative position $d = 0$.

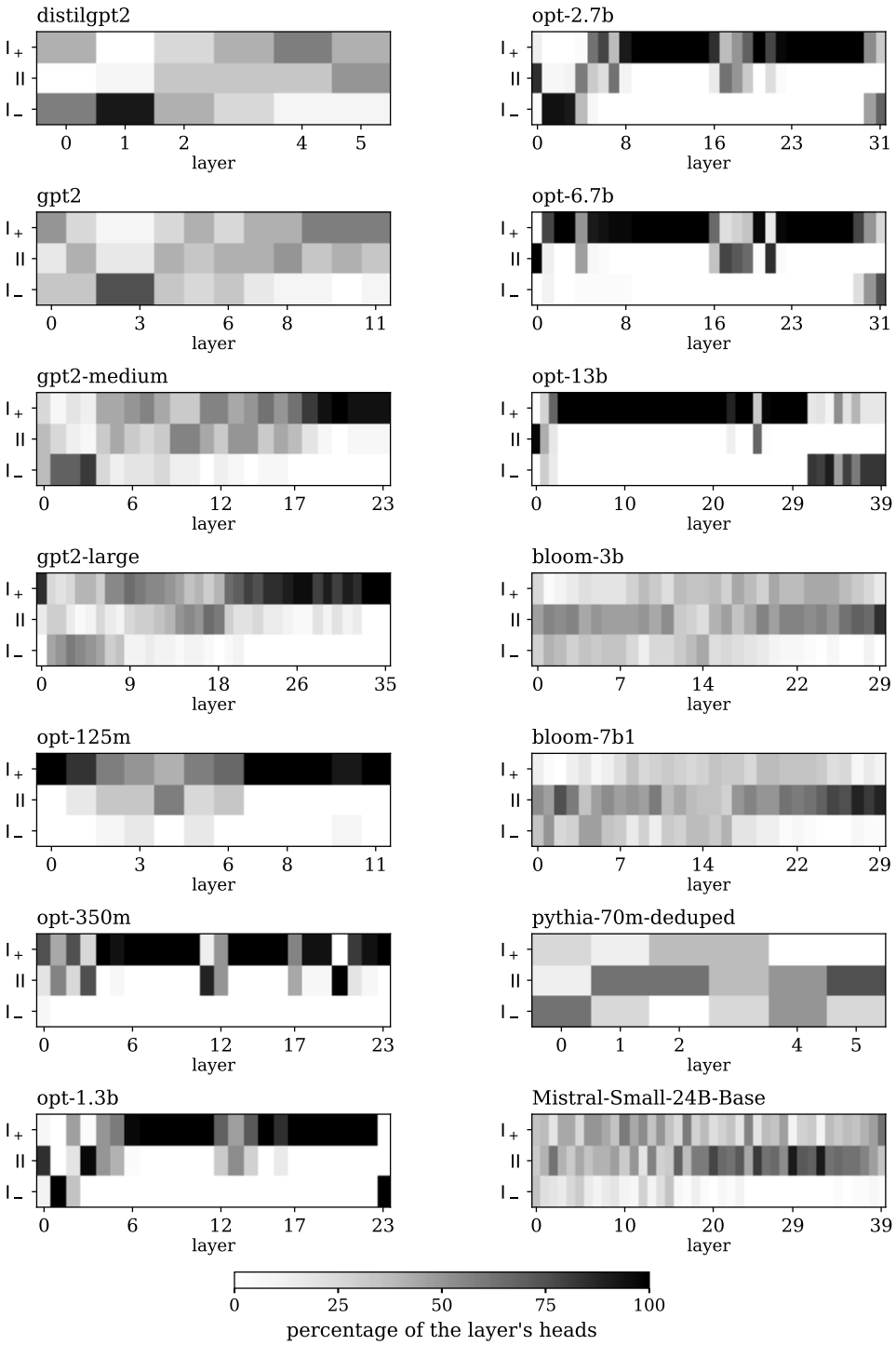

Figure 9: Parity by layer in all fourteen models. Each cell shows the percentage of a layer's heads of the indicated type, with a shared scale from white (0%) to black (100%). Types use the symmetric profile: Type II has $c_S + d_S \leq 1/50$; otherwise $c_S > d_S$ gives Type I_+ and the remaining heads are Type I_-. Layers are indexed from zero; RoPE models use relative position zero.